Genomics 98 (2011) 337–342

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A whole-genome analysis of premature termination codons

Elizabeth T. Cirulli [a], Erin L. Heinzen [a], Fred S. Dietrich [b], Kevin V. Shianna [a], Abanish Singh [a], Jessica M. Maia [a], James J. Goedert [c], David B. Goldstein [a,*]

[a] Center for Human Genome Variation, Duke University School of Medicine, Box 91009, Durham, 27708, USA
[b] Department of Molecular Genetics and Microbiology, Duke University Medical Center, Research Drive, Durham, NC 27710, USA
[c] Infections & Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, US National Cancer Institutes of Health, 6120 Executive Boulevard, Rockville, 20852, USA

ARTICLE INFO

Article history:
Received 14 April 2011
Accepted 14 July 2011
Available online 22 July 2011

Keywords:
Nonsense-mediated decay
Whole-genome sequencing
RNA-Seq
Premature termination codons

ABSTRACT

We sequenced the genomes of ten unrelated individuals and identified heterozygous stop codon-gain variants in protein-coding genes: we then sequenced their transcriptomes and assessed the expression levels of the stop codon-gain alleles. An ANOVA showed statistically significant differences between their expression levels ($p = 4 \times 10^{-16}$). This difference was almost entirely accounted for by whether the stop codon-gain variant had a second, non-protein-truncating function in or near an alternate transcript: stop codon-gains without alternate functions were generally not found in the cDNA ($p = 3 \times 10^{-5}$). Additionally, stop codon-gain variants in two intronless genes were not expressed, an unexpected outcome given previous studies. In this study, stop codon-gain variants were either well expressed in all individuals or were never expressed. Our finding that stop codon-gain variants were generally expressed only when they had an alternate function suggests that most naturally occurring stop codon-gain variants in protein-coding genes are either not transcribed or have their transcripts destroyed.

1. Introduction

Premature termination codons (PTCs) are usually destroyed through the process of nonsense-mediated decay (NMD). Without this regulation, the incorrect transcripts could be translated into proteins with potentially deleterious functions; conversely, in cases where the resulting proteins have partial or nearly normal function, it would be beneficial if NMD did not occur. Indeed, NMD (and, less commonly, its absence) has been shown to play a role in numerous diseases, including β-thalassemia, Marfan syndrome and cystic fibrosis (reviewed in [1,2]). In mammals, NMD seems to rely on the splicing and translation machinery; it can therefore only occur in protein-coding transcripts that contain introns [3,4]. Furthermore, NMD does not appear to function if the PTC is less than 50–55 bp upstream of the final splice junction of the gene (the "50 nt rule"); thus PTCs within the final exon do not tend to trigger NMD [5]. Additionally, there is some evidence that PTCs near the beginning of the transcript also do not trigger NMD [6,7]; these findings have been corroborated by genome sequencing studies that have observed that naturally-occurring PTCs cluster in the 5′ and 3′ ends of genes [8]. The NMD machinery is not entirely understood, but it seems that in addition to disposing of potentially dangerous transcripts, it also plays a major role in transcriptional regulation and the expression of alternate splice forms (reviewed in [1,9]). Organisms with mutations that disrupt the NMD machinery are often inviable [10–14].

Although there is a large body of knowledge about NMD, and it has been found to exist in all eukaryotic species examined to date, it has not previously been possible to investigate the expression of all naturally-occurring stop codon-gain variants in a genome at once. The recent advent of affordable Next-generation sequencing (NGS) technologies has now made such studies possible. Here, we have sequenced the genomes and transcriptomes of ten individuals; to assess the expression of all naturally occurring PTCs, we have identified high-quality heterozygous stop codon-gain variants in the genomes and looked for allelic imbalances for these variants in the transcriptomes.

2. Results

2.1. Identification of variants

The genomes of ten individuals were sequenced to an average coverage of 28.9×, with an average of 97% of the alignable bases of the genome (i.e., excluding gaps in the reference) covered to at least 5×. To obtain a reliable set of variants, we concentrated on high-quality coding single nucleotide variants (SNVs) in protein-coding genes that were observed in more than one of these 10 genomes, 65 control genomes, 180 control exomes or dbSNP. Because we wanted to identify allelic imbalances in the corresponding cDNA, we further restricted to heterozygous variant calls: a total of 89,435 such calls

* Corresponding author at: Center for Human Genome Variation, Duke University School of Medicine, Box 91009, Durham, NC 27708 USA. Fax: +1 919 668 6787.
E-mail address: d.goldstein@duke.edu (D.B. Goldstein).

0888-7543/$ – see front matter
doi:10.1016/j.ygeno.2011.07.001

were made in these ten genomes, corresponding to 39,112 variants as some variants were found in multiple individuals. We then searched for these variants in the PBMC transcriptome sequences from these same ten individuals. To lessen the effect of a small sample size of reads on allelic imbalances, all positions with less than 10× coverage by RNA-Seq were discarded; the remaining average of 24 MB per sample (covering approximately 34% of exonic bases) were sequenced to an average coverage of 67.0×. Thirty percent of the locations corresponding to gDNA coding variant calls had at least 10× coverage in the cDNA; 93% of these locations had a variant call that matched the gDNA call (Table 1, Fig. 1).

2.2. Imbalances at stop codon-gain variants

There were 85 heterozygous stop codon-gain variant calls, corresponding to 42 variants, matching locations with at least 10× coverage by RNA-Seq; this was a smaller portion of covered regions than for other variant types, which is not unexpected as genes heterozygous for PTC variants would conventionally be expected to be expressed at approximately half the level of normal transcripts. In contrast to the trends for stop codon-loss, nonsynonymous and synonymous variants, where 80–95% of the variants were found by RNA-Seq, only 60% of the stop codon-gains were found (Table 1). We then looked at the percent of variant reads for each variant (number of variant reads/(number of variant reads + number of reference reads)). We note that the percent of variant reads for stop codon-gain alleles in the RNA-Seq data was completely uncorrelated with the corresponding values for the gDNA, indicating that any imbalances seen were unlikely to be due to systematic biases. An examination of the 21 stop codon-gain variants that were present in more than one of the ten samples in this dataset and covered by at least 10 RNA-Seq reads revealed that they were either not expressed in anyone ($n=7$) or were expressed to some level in everyone (between 20 and 85% of variant reads; $n=14$). Indeed, an ANOVA confirmed the differences in the percent of variant reads between the variants as significant ($p=4\times10^{-16}$); if restricted to the expressed stop codon-gain variants, no differences between the variants were found ($p=0.22$). Additionally, an ANOVA failed to find a statistically significant difference in the percent of variant reads between the ten individuals ($p=0.97$), even when restricting analysis to the expressed stop codon-gain variants ($p=0.29$).

A full examination of the annotation for all 42 stop codon-gain variants showed that 13 (52%) of the 25 stop codon-gain variants that were expressed were also annotated as either nonsynonymous, synonymous, or UTR variants in alternate transcripts. For example, chr6:53,241,923C>A (G>T on the reverse strand) lies within two transcripts: it changes a GGA codon to a TGA in ENST00000370918, resulting in a PTC, and it changes a CTG codon to a CTT in ENST00000304434, resulting in a synonymous change. These are alternative transcripts for the same gene (ELOVL5); in ENST00000370918, an exon is skipped that usually transitions the phase from 1 to 0, resulting in a frameshift in the following exons that does not produce a PTC unless this variant is introduced. Zero of the 17 stop codon-gain variants that were not expressed had alternate annotations as nonsynonymous, synonymous, or UTR. Furthermore, eight stop codon-gain variants that were expressed (67% of those remaining) had received secondary annotations as intronic or upstream or downstream of a gene; this was only true of three (18%) of the stop codon-gain variants that were not expressed. A Fisher's exact test confirmed that expressed stop codon-gain variants were significantly more likely to have secondary annotations than were those that were not expressed ($p=3\times10^{-5}$).

There were only four stop codon-gain variants that were expressed but did not have alternate annotations in NCBI 36 Ensembl release 50. Further examination of these variants found that three were annotated in the updated GRCh 37 Ensembl release 57 as having additional functions in alternate transcripts. The fourth, rs2272754 in *ZC3H3*, lies only within one known transcript.

2.3. Positions and codon types of stop codon-gain variants

Thirty-one (74%) of the 42 stop codon-gain variants in this dataset were found in codon position 1: 16 (52%) of these were expressed. Although there was no statistically significant difference between the codon positions of the stop codon-gain alleles that were and were not expressed, there was a trend for the stop codon-gains in positions 2 and 3 to be expressed. Of the four variants in position 2, all were expressed, and of the seven variants in position 3, five (71%) were expressed.

There was no significant difference between the type of stop codon gained and whether or not it was expressed. Twenty-four percent of the 42 stop codon-gain variants were TAA (70% expressed), 36% were TAG (53% expressed), and 40% were TGA (59% expressed).

We also examined the relationship between the percent of variant reads and the position of the variant within the coding sequence of the gene and within the exon (Fig. 2). There was not a significant association between either of these measurements and either the percent of variant reads or whether the variant was expressed. Additionally, we compared the percent of variant reads for stop codon-gains that were found in the first ($n=7$; 17%), the last ($n=14$; 33%), a middle ($n=19$; 45%), or the only ($n=2$; 5%) exon of the gene (Fig. 2). Again, there was not a statistically significant difference between these groups, but there was a trend for stop codon-gains found in the last exon of the gene to have higher expression ($p=0.09$). Finally, we compared the percent of variant reads of stop codon-gains found less than 55 bp upstream of the last splice junction ($n=17$; 40%) to those farther upstream (Fig. 2): again there was no statistically significant difference, but there was a trend for higher expression of those less than 55 bp upstream of the last splice junction ($p=0.10$).

2.4. Expression of genes containing stop codon-gain variants

For individuals with the same stop codon-gain variant, there was no statistically significant difference between the average reads per kb of exon model per million mapped reads (RPKM) of genes with expressed stop codon-gains and those with stop codon-gains that were not expressed. There was a trend ($p=0.07$) for those genes with stop codon-gains that were not expressed to have a higher RPKM, but this was found to be driven by two outliers: stop codon-gains that

Table 1
Data for cDNA positions corresponding to the locations of gDNA exonic variants. All cDNA positions were required to have a coverage of at least 10×. The mean % of variant reads for covered corresponding cDNA positions includes positions where no cDNA variant was called.

	Stop codon-gain	Stop codon-loss	Non-synonymous	Synonymous	All coding variants
High-quality gDNA variant calls	503	96	43,273	45,563	89,435
Corresponding covered cDNA positions	85	20	10,942	16,006	27,053
(% of gDNA calls)	(16.9%)	(20.8%)	(25.3%)	(35.1%)	(30.2%)
Corresponding covered cDNA positions with variant called	51	16	9,977	15,158	25,202
(% of covered corresponding positions)	(60.0%)	(80.0%)	(91.2%)	(94.7%)	(93.2%)
Mean % of variant reads for covered corresponding cDNA positions	28.0%	41.7%	46.2%	47.4%	46.9%

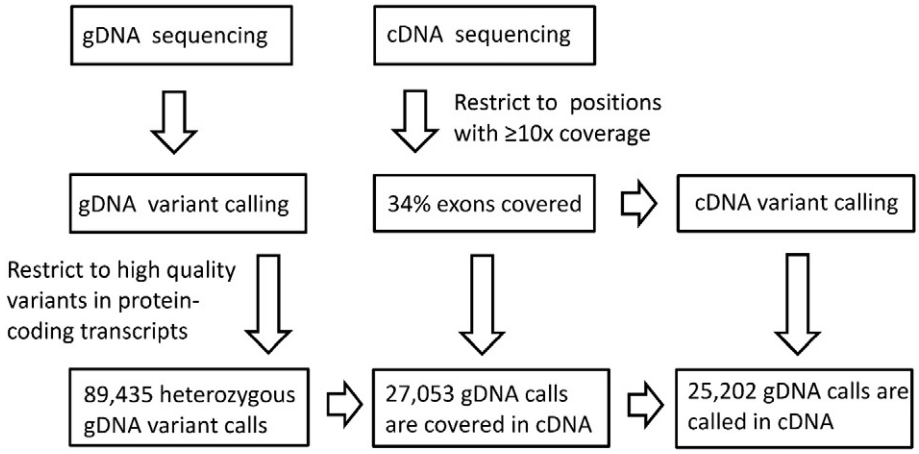

Fig. 1. Flowchart of methods.

were not expressed but lay within genes with very high transcription levels. These two genes were also found to have high RPKMs in individuals who did not have the stop codon-gain variants.

For each gene, the average RPKM for individuals with a stop codon-gain was also compared to the average RPKM for individuals without stop codon-gains. A paired t-test of these average RPKMs did not reveal a significant difference between the groups.

2.5. Unique stop codon-gain variants

Although the above analyses focused on more common variants, we also examined unique stop codon-gain variants that were observed in only one of the ten samples and none of the 245 controls or dbSNP. There were 32 such variants in the gDNA; similar to what was observed for more common variants, 5 (16%) were covered by at least 10 reads in the cDNA. The stop codon-gain allele was not expressed for three of these five variants; the other two had 42% and 71% of variant reads. One of the expressed stop codon-gains was found to have alternate annotations as both a synonymous variant and a 5′UTR variant, while the stop codon-gains that were not expressed had no alternate annotations. Additional analysis of the other expressed stop codon-gain variant showed that it was included in a non-coding transcript in GRCh 37 Ensembl release 57.

The properties of the RPKM values, codon types, and positions of the unique stop codon-gain variants were not significantly different

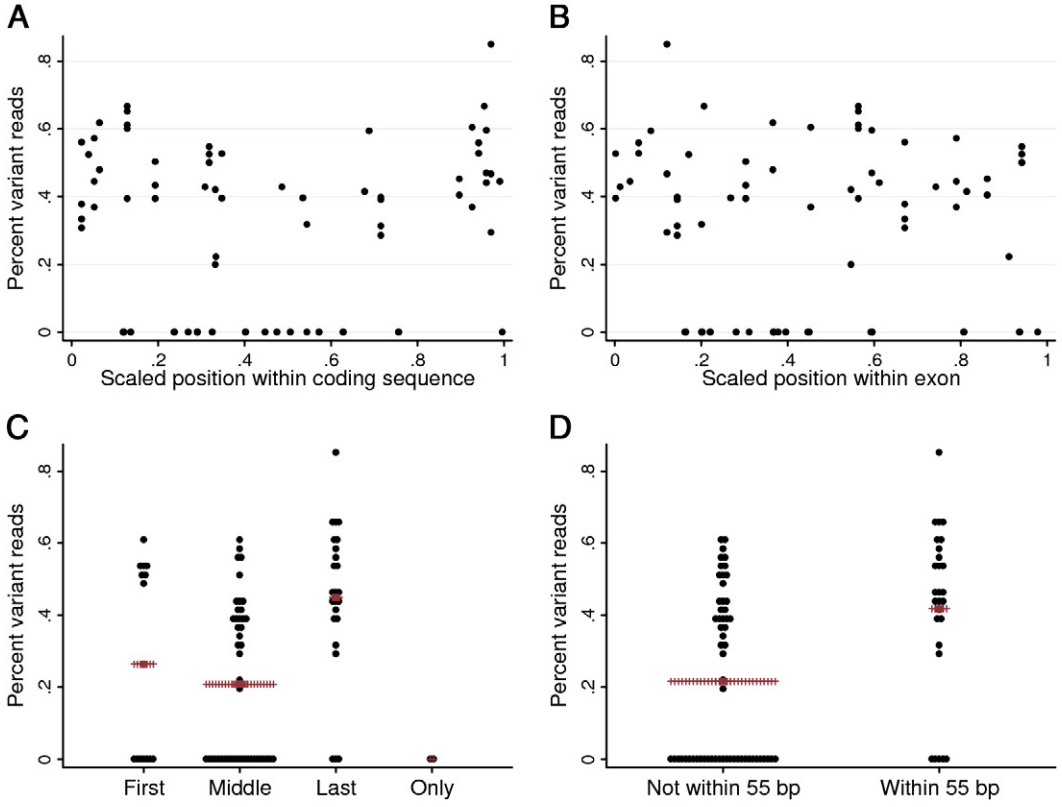

Fig. 2. Percent stop codon-gain variant reads. Shown by scaled position within a) the coding sequence and b) the exon (which in some cases contains UTR). The values shown are scaled from 0 to 1 and are per individual, not averaged across individuals. The percent of variant reads are also shown broken down by c) exon ("Only" indicates that the gene has no introns) and by d) whether the variant is within 55 bp of the final splice junction (this category necessarily includes the entire final exon of each gene). Red crosses in c and d indicate the means for each group.

from those of the more common variants. However, there was a trend ($p = 0.08$) for unique stop codon-gains to occur closer to the 3′ end of the exon than more common variants.

3. Discussion

In this study, we have examined the effect of PTCs on gene expression by studying allelic imbalances at heterozygous stop codon-gain variants in the RNA-Seq data of ten unrelated individuals. We have found that the expression of alleles containing stop codon-gain variants is generally blocked; the expression of such alleles largely appears to occur only when the variant is also part of another transcript, where it does not encode a stop codon-gain.

We feel that the most likely explanation for the imbalances seen in the cDNA is NMD. It is worth noting, however, that there is potential for the imbalances to be due to other factors, such as unequal transcription. One possible explanation for the lack of expression of some stop codon-gain variants would be that they lie within an imprinted gene. However, none of the genes with stop codon-gain variants studied here are found in the Catalogue of Parent of Origin Effects or Geneimprint, although the lists provided by these sources are of course not completely comprehensive [15,16]. Another possible explanation could be RNA editing [17,18]. Additionally, NGS is not without errors, and a false positive variant call in the gDNA could lead to what seems like an unexpressed variant in the cDNA. Although we focused on high-quality gDNA variant calls, we did not confirm the variants studied here through additional genotyping methods. It must also be noted that our RNA-Seq data was based solely on RNA from PBMCs, so some of the imbalances seen here may be tissue specific.

The results of this study demonstrate the importance of remembering that variants can often have multiple annotations. Twenty-four (57%) of the 42 stop codon-gains examined here had alternate annotations; of these 24, 21 were expressed. Although some of these alternate annotations were not actually within a transcript (e.g., intronic or upstream), it seems possible that variants with multiple annotations may lie within alternate transcripts that have not yet been described or within transcripts that have been incorrectly annotated as coding. For example, chr19: 63785276C>T is predicted to be a stop codon-gain in ENST00000312426 and an intronic variant in ENST00000237694 in NCBI 36 Ensembl release 50, but in GRCh 37 Ensembl release 57, it is annotated as a variant in the non-coding transcript ENST00000493504. Conversely, of the 25 expressed stop codon-gain variants, only 4 did not have alternate annotations. Further investigation revealed that three of these four had alternate annotations in the updated GRCh 37 Ensembl release 57, leaving only one stop codon-gain variant that was both expressed and lacking a secondary annotation in or near an alternate transcript. It is possible that this variant has a secondary function in an alternate transcript that has not yet been described; however, examination of the expression patterns in this region did not suggest the existence of another transcript. An alternative explanation is that this variant is expressed due to its position proximal to the C terminus (position 0.93 in the coding sequence); previous studies have shown that such PTCs are often expressed (reviewed in [9]). Indeed, 28% of the 25 stop codon-gain variants that were expressed had scaled coding positions of 0.9 or greater; this was only true of 6% of the 17 stop codon-gain variants that were not expressed (Fig. 2). Although there was no significant association between the scaled positions within the coding sequence of each PTC and either the average percent of variant reads or whether a variant was called in the cDNA, there was a significant trend for stop codon-gain variants with secondary exonic annotations to be closer to the C termini of genes ($p = 0.007$).

We also examined whether variants in this dataset followed the known "rules" of NMD: that for NMD to be triggered, the protein-coding gene must contain introns and the variant must be greater than 55 coding bases from the last exon junction of the gene [3–5]. Only two genes with stop codon-gain variants, *PABPC3* and *AL592309.24* (HS6ST1P1), were without introns (including the UTRs). Contrary to expectation, the stop codon-gain alleles in both of these intronless genes were not expressed. It should also be noted that *AL592309.24* is now annotated in GRCh 37 Ensembl release 57 as a processed pseudogene, which are also not expected to undergo NMD. We also found that 48% of the 25 stop codon-gains that were expressed were found less than 55 bp upstream of the final splice junction; however, this was also true of 29% of those that were not expressed (Fig. 2). While these results do not support the 50 nt rule, there is prior evidence that this rule is not well followed in humans [19].

We did not find a significant difference between the expression levels of genes with expressed and unexpressed stop codon-gains. Furthermore, we did not find a significant difference between the expression levels of those individuals who had stop codon-gains and those that did not. This is surprising given that genes not expressing one allele would be expected to have lower transcript levels. While past studies have shown that the overall expression of genes decreases when they are undergoing NMD (for example, see [20]), these studies have tended to focus on stop codon-gains that are associated with diseases. A large-scale analysis of the effects of stop codon-gains that are not associated with diseases on total expression levels has not yet been performed. Although our study was performed on a small number of samples and requires replication, our results suggest that when seemingly harmless stop codon-gains are not expressed, the loss of transcript may be compensated for with increased expression of the alternate allele, perhaps via a feedback mechanism. This observation implies that many heterozygous stop codon-gain variants may not actually produce functional consequences. If this is true, then it is a point worth considering in the interpretation of sequencing studies, where stop codon-gain variants are often categorized as being of the highest importance.

In this paper, we focused on SNVs, although PTCs can also occur as a consequence of a small indel or larger structural variant. We chose to only examine SNVs in this study because with the current technology, this type of variant call is by far the most reliable [2]. In addition to the fact that there are many false positive structural variant and indel calls, these types of variants would also not be able to be assessed directly in the RNA-Seq data: it is not possible to use such data to call structural variants by the read depth method, and the calling of indels is difficult because of the need to align reads that cross splice junctions. As another means of assuring that we focused on high-quality variants, in the main analyses of this paper we only examined variants that were found in more than one individual (including 245 controls and dbSNP). However, we also examined unique stop codon-gain variants separately and found that their properties were not noticeably different from those of the more common stop codon-gain variants.

Previous studies have demonstrated inter-individual differences in NMD efficiency and differences in the propensity of different PTCs to succumb to NMD (reviewed in [9]). However, in this study we have found that naturally occurring stop codon-gain variants are generally not expressed unless they have secondary annotations in or near other transcripts. Though our sample size is admittedly small, we do not find inter-individual or inter-variant differences in PTC expression.

While there have been numerous studies of NMD, ours is the first study to take an unbiased look at PTCs across a fully sequenced genome and examine their level of expression. Such an analysis was not possible until very recently; before Next Generation sequencing technologies arose, identifying all of the stop codon-gain variants that were present in multiple genomes would not have been feasible. The results of our study suggest that NMD, which has long been known to work efficiently in the specific genes in which it has been studied,

does indeed work well even when investigated in an unbiased, whole-genome manner.

4. Methods

4.1. Samples, preparation and sequencing

The samples used in this study represent a partial overlap with those presented in previous works; six of the ten genomes discussed have already been published [2], as has one of the transcriptomes [21].

DNA was extracted from peripheral blood mononuclear cells (PBMCs) using the QIAGEN Autopure LS. RNA was extracted from viable PBMCs using the Qiagen RNeasy kit. The DNA was prepared for sequencing according to Illumina's gDNA sample prep kit protocol: briefly, the DNA is randomly fragmented by nebulization, then undergoes end repair, a single A base is added, adaptor ligation occurs, a gel is run to isolate appropriately-sized fragments, and finally polymerase chain reaction (PCR) amplification is performed. One microgram of total RNA was prepared for each sample according to the Illumina RNA-seq protocol: briefly, globin reduction, polyA enrichment, chemical fragmentation of the polyA RNA, cDNA synthesis, and size selection of the cDNA products. Next, the size-selected libraries were used for cluster generation on the flow cell. All prepared flow cells were run on the Genome Analyzer II using the paired-end module.

4.2. Alignment

gDNA was aligned to the reference genome (NCBI Build 36) using the BWA software [22]. cDNA was aligned to the reference genome using TopHat [23]. The -GFF option utilized a transcript library downloaded from Ensembl release 50 to specify known protein-coding transcripts and splice junctions (although novel junctions were permitted), and the library was screened to remove contigs and mitochondrial DNA. The mate inner distance (option –r) was calculated using the fragment and read sizes of each sample. A minimum anchor length (option –a) of four was used. To assist in alignment to small exons, the 100 bp reads were broken down into four segments of default length 25 bp that were then joined back together after being individually aligned. Two mismatches were permitted per 25 bp segment, and no mismatches were permitted in the 4 bp anchor region on either side of a splice junction. Introns were permitted to range in size from 10 bp to 500 kb.

4.3. Transcript expression quantitation and variant identification

RPKM values were calculated using Partek software and visualized using custom tracks in the UCSC genome browser [24].

SAMtools was used to call SNVs in the gDNA and cDNA alignments. Indels were not considered because they are less reliable [2] and the TopHat [23] alignment did not support the gaps required for indel calls. Prior to variant calling, SAMtools first removed potential PCR duplicates via the rmdup (paired reads) command [25]. It was also used for SNV identification, using the pileup command with the –c option and default settings. The SNVs were then filtered using SAMtools' variation filter with the default settings but changing the filter for the maximum allowed coverage per variant to 254 for gDNA and 1 million for cDNA. SNVs lying outside exons as defined by the transcript library were removed.

Structural variants (SVs) were identified from the whole-genome sequences using ERDS v.1.02 (http://www.duke.edu/~mz34/erds.htm) with a sliding window size of 2 kb. None of the expressed stop codon-gain variant calls in this paper were found within regions predicted to be duplicated by this method.

4.4. Annotation, QC and visualization

The Sequence Variant Analyzer (SVA) software (svaproject.org) [26] was used to annotate the function (synonymous, nonsynonymous, stop codon-gain or stop codon-loss) of each variant according to NCBI 36 Ensembl release 50. After annotation, SVA was used to exclude all gDNA SNVs that did not meet minimum QC requirements: a SNP quality score of at least 20, a consensus score of at least 20, and at least 3 reads supporting the variant. cDNA SNVs called by SAMtools were not further screened by SVA. Analysis was restricted to those variants lying within the 21,021 genes that were annotated as protein-coding. Furthermore, we excluded all variants that exhibited allelic imbalances in the gDNA, requiring that the percent of variant reads (number of variant reads/(number of variant reads + number of reference reads)) was between 40 and 60%. Additionally, we restricted our analysis to the autosomes as all subjects were male. For the main analysis, we focused only on variants that were present in more than one individual, considering the other nine samples as well as dbSNP, 65 control genomes and 180 control exomes that had been sequenced as part of other projects. This was done to increase the likelihood that the SNVs were real. SVA was used to visualize the variants and to determine the position within the coding sequence of stop codon-gain variants; all positions were scaled so that 0 represented the first base and 1 represented the last.

4.5. Identification of allelic imbalances and statistical analysis

For the RNA-Seq data, analysis was restricted to locations with at least 10× coverage. The RNA-Seq percent of variant reads was determined for each variant, and locations that had no variant called were assumed to have 0% of variant reads. The number of gDNA-matching variants called by RNA-Seq and the average percent of variant reads can be found in Table 1.

All statistical analyses were performed using STATA with a p-value cutoff of 0.05 [27]. Comparisons between two continuous variables were made using a linear regression, comparisons between a continuous variable and a discrete variable were made using ANOVA, and comparisons between two discrete variables were made using Fisher's exact test. A paired t-test was used to compare each gene's average expression in individuals with a stop codon-gain to those without.

Acknowledgments

Funding was provided by the NIAID Center for HIV/AIDS Vaccine Immunology grant UO1AI067854 and the Bill & Melinda Gates Foundation. We also acknowledge C. Gumbs, K. Cronin and L. Little for DNA and RNA extraction. Sequencing of control genomes and exomes was funded by NIMH GrantRC2MH089915 and NINDS Award# RC2NS070344, and the control samples were provided by G Cavalleri, S Sisodiya, C Depondt, R Radtke, A Husain, M Mikati, N Walley, JP McEvoy, AC Need, J Silver, M Silver, and R Ottman.

References

[1] P. Nicholson, H. Yepiskoposyan, S. Metze, R. Zamudio Orozco, N. Kleinschmidt, O. Muhlemann, Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors, Cell Mol. Life Sci. 67 (2010) 677–700.

[2] K. Pelak, K.V. Shianna, D. Ge, J.M. Maia, M. Zhu, J.P. Smith, E.T. Cirulli, J. Fellay, S.P. Dickson, C.E. Gumbs, E.L. Heinzen, A.C. Need, E.K. Ruzzo, A. Singh, C.R. Campbell, L.K. Hong, K.A. Lornsen, A.M. McKenzie, N.L. Sobreira, J.E. Hoover-Fong, J.D. Milner, R. Ottman, B.F. Haynes, J.J. Goedert, D.B. Goldstein, The characterization of twenty sequenced human genomes, PLoS Genet. 6 (2010).

[3] R. Thermann, G. Neu-Yilik, A. Deters, U. Frede, K. Wehr, C. Hagemeier, M.W. Hentze, A.E. Kulozik, Binary specification of nonsense codons by splicing and cytoplasmic translation, EMBO J. 17 (1998) 3484–3494.

[4] K.S. Brocke, G. Neu-Yilik, N.H. Gehring, M.W. Hentze, A.E. Kulozik, The human intronless melanocortin 4-receptor gene is NMD insensitive, Hum. Mol. Genet. 11 (2002) 331–335.
[5] E. Nagy, L.E. Maquat, A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance, Trends Biochem. Sci. 23 (1998) 198–199.
[6] R. Asselta, S. Duga, S. Spena, E. Santagostino, F. Peyvandi, G. Piseddu, R. Targhetta, M. Malcovati, P.M. Mannucci, M.L. Tenchini, Congenital afibrinogenemia: mutations leading to premature termination codons in fibrinogen A alpha-chain gene are not associated with the decay of the mutant mRNAs, Blood 98 (2001) 3685–3692.
[7] L.W. Harries, C. Bingham, C. Bellanne-Chantelot, A.T. Hattersley, S. Ellard, The position of premature termination codons in the hepatocyte nuclear factor −1 beta gene determines susceptibility to nonsense-mediated decay, Hum. Genet. 118 (2005) 214–224.
[8] P.C. Ng, S. Levy, J. Huang, T.B. Stockwell, B.P. Walenz, K. Li, N. Axelrod, D.A. Busam, R.L. Strausberg, J.C. Venter, Genetic variation in an individual human exome, PLoS Genet. 4 (2008) e1000160.
[9] G. Neu-Yilik, A.E. Kulozik, NMD: multitasking between mRNA surveillance and modulation of gene expression, Adv. Genet. 62 (2008) 185–243.
[10] S.M. Medghalchi, P.A. Frischmeyer, J.T. Mendell, A.G. Kelly, A.M. Lawler, H.C. Dietz, Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability, Hum. Mol. Genet. 10 (2001) 99–105.
[11] M.M. Metzstein, M.A. Krasnow, Functions of the nonsense-mediated mRNA decay pathway in Drosophila development, PLoS Genet. 2 (2006) e180.
[12] J. Weischenfeldt, I. Damgaard, D. Bryder, K. Theilgaard-Monch, L.A. Thoren, F.C. Nielsen, S.E. Jacobsen, C. Nerlov, B.T. Porse, NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements, Genes Dev. 22 (2008) 1381–1396.
[13] N. Wittkopp, E. Huntzinger, C. Weiler, J. Sauliere, S. Schmidt, M. Sonawane, E. Izaurralde, Nonsense-mediated mRNA decay effectors are essential for zebrafish embryonic development and survival, Mol. Cell. Biol. 29 (2009) 3517–3528.
[14] M. Yoine, T. Nishii, K. Nakamura, Arabidopsis UPF1 RNA helicase for nonsense-mediated mRNA decay is involved in seed size control and is essential for growth, Plant Cell Physiol. 47 (2006) 572–580.
[15] Geneimprintin:, www.geneimprint.com.
[16] Catalogue of Parent of Origin Effectsin:, www.otago.ac.nz/IGC.
[17] S.P. Shah, R.D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R.A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G.A. Taylor, A.E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R.L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M.A. Marra, S. Aparicio, Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, Nature 461 (2009) 809–813.
[18] I. Chepelev, G. Wei, Q. Tang, K. Zhao, Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq, Nucleic Acids Res. 37 (2009) e106.
[19] G. Neu-Yilik, N.H. Gehring, R. Thermann, U. Frede, M.W. Hentze, A.E. Kulozik, Splicing and 3′ end formation in the definition of nonsense-mediated decay-competent human beta-globin mRNPs, EMBO J. 20 (2001) 532–540.
[20] L. Linde, S. Boelz, M. Nissim-Rafinia, Y.S. Oren, M. Wilschanski, Y. Yaacov, D. Virgilis, G. Neu-Yilik, A.E. Kulozik, E. Kerem, B. Kerem, Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin, J. Clin. Investig. 117 (2007) 683–692.
[21] E.T. Cirulli, A. Singh, K.V. Shianna, D. Ge, J.P. Smith, J.M. Maia, E.L. Heinzen, J.J. Goedert, D.B. Goldstein, Screening the human exome: a comparison of whole genome and whole transcriptome sequencing, Genome Biol. 11 (2010) R57.
[22] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics (Oxford, England) 25 (2009) 1754–1760.
[23] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.
[24] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[25] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[26] D. Ge, E.K. Ruzzo, K.V. Shianna, M. He, K. Pelak, E.L. Heinzen, A.C. Need, E.T. Cirulli, J.M. Maia, S.P. Dickson, M. Zhu, A. Singh, A.S. Allen, D.B. Goldstein, SVA: Software for Annotating and Visualizing Sequenced Human Genomes, Bioinformatics 27 (14) (2011) 1998–2000.
[27] StatCorp, Intercooled stata, Intercooled Stata 9.2 for Windows, College Station, TX 77845 USA, 2006.